**HW06**

Use the GFCLOCKS data (Ex. 8.30 8th edition, Ex. 8.26 7th edition) found here  [GFCLOCKS.csvPreview the document](https://rutgers.instructure.com/courses/39544/files/8197143/download?wrap=1)

and the code below to generate the diagnostic plots listed below.  **Summarize the key findings regarding outliers.**

##HW06  
library(faraway) #this command brings in a library of regression functions  
library(psych)  
library(olsrr)  
library(car)

#Read in the data set GFCLOCKS.csv  
GFCLOCKS <- read.csv(file="C:/Users/jmard/Desktop/RegressionMethodsSpring2020/Homework/GFCLOCKS.csv",header = TRUE)

head(GFCLOCKS,5L)

lmod<-lm(PRICE ~ AGE + NUMBIDS, data=GFCLOCKS)

ols\_plot\_resid\_fit(lmod)

spreadLevelPlot(lmod,robust.line=FALSE,grid=TRUE,smooth=TRUE)

ols\_plot\_resid\_stud(lmod)

ols\_plot\_resid\_stand(lmod)

ols\_plot\_resid\_lev(lmod)

ols\_plot\_resid\_stud\_fit(lmod)

**What you need to submit into Canvas for this assignment in order of priority:**

**1.  A paragraph or two summarizing your findings.**

**2.  At least one plot from those in the list.**

**HW07**

Use the GFCLOCKS data set from HW06 to generate the following plots:

ols\_plot\_cooksd\_chart(lmod)

ols\_plot\_dfbetas(lmod)

ols\_plot\_dffits(lmod)

ols\_plot\_hadi(lmod)

ols\_plot\_resid\_plot(lmod)

**Summarize your findings in terms of leverage and influence.**

**What you need to submit into Canvas for this assignment in order of priority:**

**1.  A paragraph or two summarizing your findings.**

**2.  At least one plot from those in the list.**

**HW08**  
Use the presidents data set ([presidents.csvPreview the document](https://rutgers.instructure.com/courses/39544/files/8436673/download?wrap=1)) that shows quarterly approval ratings of US presidents  
during 30 years starting in 1945.

a) Generate two lowess plots using spans that you choose between 0.05 and 1.  
b) Provide 1 or 2 sentences describing the pattern for the span you choose.  
  
c) Generate two smoothing plots with normal kernel density standard deviation h between .2 and .8  
d) Provide 1 or 2 sentences describing the pattern for the h you choose.

#See useful R code below.

#read in the data which is in a csv file  
presidents <- read.csv(file="Your directory where the presidents.csv data set is stored/presidents.csv",header = TRUE)  
head(presidents)  
str(presidents)

library(faraway)  
#code to use for different spans. The one below uses a span of 0.20  
with(presidents,{  
plot(presidents ~ quarter, col=gray(0.1))  
f <- loess(presidents ~ quarter,span=0.20)  
i <- order(quarter)  
lines(f$x[i],f$fitted[i])  
})

#code to use for different choices of h. The one below uses an h  
library(sm) #install if not already installed  
with(presidents,sm.regression(x=quarter,y=presidents,h=.1))

**HW09**

The National Institute of Diabetes and Digestive and Kidney Diseases

conducted a study on 768 adult female Pima Indians living near Phoenix.

The pima dataset resulting from the study is available in R and contains the following variables:

**test** - results of a test to determine if the female patient shows signs of diabetes

          (coded 0 if negative, 1 if positive)

**age** - Age (years)

**bmi** - Body mass index (weight in kg/(height in metres squared))

**diastolic** - Diastolic blood pressure (mm Hg)

**diabetes** - Diabetes pedigree function

**glucose** - Plasma glucose concentration at 2 hours in an oral glucose tolerance test

**insulin** - 2-Hour serum insulin (mu U/ml)

**pregnant** - Number of times pregnant

**triceps** - Triceps skin fold thickness (mm)

The following logistic model with test as the response was fit with the following table results:

logistic <- glm(test ~ age + bmi + diastolic + diabetes + glucose + insulin + pregnant +

                                 triceps,family=binomial(logit),data=pima)

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | **Reference** | |
|  |  | **0** | **1** |
| **Prediction** | **0** | **446** | **117** |
| **1** | **54** | **151** |

A positive test result (test=1) was modeled as an event.  Compute the following measures:

Accuracy

Sensitivity

Specificity

Positive Predictive Value

Negative Predictive Value

**HW10**

See HW09 for details regarding the pima data set.

Use the Rcode below to obtain the crude and adjusted Odds Ratios.

1. a) Interpret the adjusted Odds Ratio for diastolic.
2. b) Perform the Likelihood Ratio Test to determine if age, insulin, and triceps can be dropped from the model using alpha=0.05.
3. c) Perform the Likelihood Ratio Test to determine if the coefficient of diastolic is equal to the coefficient of insulin using alpha=0.05.

library(faraway)  
library(lmtest)  
library(epiDisplay)

#Multiple Logistic Regression in R

logistic <- glm(test ~ age + bmi + diastolic + diabetes + glucose + insulin + pregnant + triceps,family=binomial(logit),data=pima)  
summary(logistic)

## Easier view  
logistic.display(logistic)

#Likelihood Ratio Test for H0: B(triceps)=0 #write model without triceps  
logistic1 <- glm(test ~ age + bmi + diastolic + diabetes + glucose + insulin + pregnant,family=binomial(logit),data=pima)  
lrtest(logistic1,logistic)

**#HW11**

#See HW09 for the description of the pima dataset.

#Use the Rcode below to perform forward, backward, and bi-directional stepwise #regression.  Provide a paragraph summarizing the results.

#Multiple Logistic Regression in R  
  
library(faraway)  
library(StepReg)

logistic <- glm(test ~ age + bmi + diastolic + diabetes + glucose + insulin + pregnant + triceps, family=binomial(logit),data=pima)  
summary(logistic)  
#analysis(logistic)

anova(logistic)

y <- "test"  
stepwiselogit(data=pima,y, exclude = NULL, include = NULL, selection = "forward",  
select = "AIC", sle = 0.15, sls = 0.15)

stepwiselogit(data=pima,y, exclude = NULL, include = NULL, selection = "backward",  
select = "AIC", sle = 0.15, sls = 0.15)

stepwiselogit(data=pima,y, exclude = NULL, include = NULL, selection = "bidirection",  
select = "AIC", sle = 0.15, sls = 0.15)

##-------------------------------------------------------------------##

**HW12**The FACTORS data set is described in Exercise 3.68 (7th edition) and Exercise 9.31 (8th edition) and is available here [FACTORS.csvPreview the document](https://rutgers.instructure.com/courses/39544/files/8636813/download?wrap=1).

#libraries needed  
library(faraway)  
library(psych)  
library(lmtest)

a) Find the mean and variance of the dependent variable y= length of stay.

b) Plot LOS versus FACTORS.

c) Fit a Poisson model of y (LOS) on the number of factors (FACTORS).

d) Determine whether the quadratic term FACTORS^2 should be added to the model.

e) What is the difference in the AICs between the model including only FACTORS in the model with the model containing both FACTORS and FACTORS^2.

Submit the plot and output and responses to d) and e) into Canvas.

**THE ONLY PLOT THAT IS NEEDED IS THE PLOT FOR b)**

**HW13**  
#The following R code generates a ridge regression using all 100 variables in the meatspec dataset in the faraway library.

library(faraway)  
library(psych)  
library(QuantPsyc)  
library(MASS)

hw13 <- meatspec  
hw13 <- data.frame(hw13)  
describe(hw13)

lmod <- lm(fat ~ .,data=hw13)  
summary(lmod)

rgmod <- lm.ridge(fat ~ .,data=hw13, lambda = seq(0, 6e-08, len=21))  
matplot(rgmod$lambda, coef(rgmod), type="l", xlab=expression(lambda),ylab=expression(hat(beta)),col=1)  
which.min(rgmod$GCV) #use the generalized cross validation (GCV) estimate  
abline(v=3.0e-08)  
rgmod$coef #prints out the coefficients

##-------------End of Program-----------------##

Use the program above as a template to generate a ridge regression on variables V71-V100.

**QUESTION:** Are the ridge regression estimates for V71-V100 at the GCV chosen lambda similar to the ridge regression estimates for V71-V100 at the GCV chosen  
lambda for the ridge regression performed on all 100 variables?

**HELPFUL HINT: #start your R program with these next 7 lines of code to select variables V71-V100.  
library(faraway)  
library(psych)  
library(QuantPsyc)  
library(MASS)  
hw13 <- meatspec[,-(1:70)]  
hw13 <- data.frame(hw13)  
describe(hw13)**

##---------------------------------------------##

**HW14**  
#The following R code performs LASSO selection using all 100 variables in the meatspec dataset in the faraway library

library(faraway)  
library(glmnet) # include glmnet library including functions for LASSO  
#------------------------------------------------#

# glmnet requires the x matrix and the response vector

hw14x <- as.matrix(meatspec[,-101]) #creates the x matrix for all 100 predictors V1-V100  
hw14y <- meatspec[, 101] #creates the y vector

lassomod <- glmnet(hw14x,hw14y)  
plot(lassomod)  
print(lassomod)

cvfit <- cv.glmnet(hw14x,hw14y)  
plot(cvfit)  
cvfit$lambda.min  
coef(cvfit,s="lambda.min") #variables with non-zero coefficients should be included in the LASSO model

##-------------End of Program-----------------##  
Use the program above as a template to generate a LASSO selection on variables V1-V30.  
**QUESTION: List the variables among V1-V30 that are selected by the LASSO selection on variables V1-V30.  This is the only information that needs to be submitted in Canvas.**

HELPFUL HINT: #start your R program with these next 2 lines of code to select variables V1-V30.  
hw14x <- as.matrix(meatspec[,1:30]) #includes the the first 30 variables V1 - V30  
hw14y <- meatspec[, 101] #creates the y vector